UNIVERSITÀ DEGLI STUDI
DI TRENTO

Centro Interdipartimentale Mente/Cervello

# Masaryk University dissertation thesis report

Study programme: Ph.D., Faculty of Informatics

Field of study: Informatics, Computational Linguistics

Name of student: Jan Pomikálek

Thesis title: Removing Boilerplate and Duplicate Content from Web Corpora

Opponent: Marco Baroni

Institution: Center for Mind/Brain Sciences, University of Trento, Italy

In natural language processing, linguistics and lexicography, it is increasingly important to have access to large amounts of text data for various forms of linguistic and statistical analysis, for knowledge extraction, and in order to train machine learning algorithms. The World Wide Web offers, to interested researchers, text on a gigantic scale. Consequently, the last decade has witnessed various attempts to construct very large corpora (large samples of text, typically pre-processed to optimize their linguistic analysis) by harvesting Web data.

Jan Pomikálek's thesis deals with -- and to a certain extent solves -- two important and often overlooked problems in the Web corpus creation pipeline, namely "cleaning" Web text from *boilerplate,* i.e., those sections of a document that are textual but not linguistically informative (advertisement, headers, navigation information, etc.), and removing duplicated text (duplicated documents, as well as sections repeated across documents). For both problems, Pomikálek proposes new algorithms, and demonstrates their efficiency as well as effectiveness.

Despite their importance, boilerplate cleaning and duplicate removal have received very little attention from the (computational) linguistics community. Pomikálek's thesis is indeed, to the best of my knowledge, the first book-length treatment of these topics, and as such I hope that it will have a strong impact on the relevant research communities. Jan Pomikálek's has already presented some of the results reported in the thesis, as well as preliminary and related work leading to the thesis, in a sizable number of publications, including articles in proceedings of high-impact conferences such as LREC, EAMT, EuraLex and CICLING. I hope that he will soon summarize the main results of his thesis work in one or more publications submitted to the leading journals in the fields of computational and corpus-based linguistics.

The thesis is divided into 4 chapters. After the introduction (chapter 1), that surveys corpora in general as well as in the Web-derived variety, chapter 2 presents the problem of boilerplate removal, current algorithms, and the novel jusText model. In an empirical evaluation on various

gold standards, jusText is shown to be competitive with state-of-the-art boilerplate removal techniques. Perhaps most interestingly, jusText is the first algorithm that fully takes context into account, when identifying blocks of good or "bad" text. In this way, it deals, to a certain extent, with the problem of fragmentation, that is, the challenge to produce, in output, sequences of coherent blocks of text, rather than detached fragments. Chapter 3 presents a survey of de-duplication algorithms, mostly developed by the information retrieval community and aimed at spotting nearly identical documents. Pomikálek shows that documents with an intermediate level of similarity, while problematic in the construction of Web corpora, are both commonly encountered in the Web and missed by algorithms searching for near-identical documents. A new algorithm is proposed that is at the same time very efficient and able to perform the more granular n-grams-based analysis that is necessary to capture intermediate levels of similarity. The algorithm is combined with jusText to allow de-duplication at the infra-document level (removing duplicate parts of documents, instead of documents as a whole), while avoiding fragmentation. The performance of the new algorithm is illustrated with large corpora in English, German and Italian. The algorithm is shown to be efficient and to capture a good proportion of duplicate blocks (although this is only demonstrated by indirect evidence, and not on a gold standard). A short conclusion (chapter 4) summarizes the main achievements of the dissertation.

Pomikálek's work is very timely, original, and methodologically and empirically sound. My overall assessment of it is very positive. However, I find that one crucial aspect is missing from the thesis, namely quantitative (or at least, extensive qualitative) evidence that boilerplate cleaning and de-duplication are indeed improving the quality of the corpora they are applied to. It should be relatively straightforward to use pre- and post-cleaned corpora to train a natural language processing algorithm (e.g., language modeling, extraction of semantic models, weakly supervised POS tagging, etc.), and then compare the results with and without cleaning on the relevant tasks. If the applied focus is on lexicography, the author could have run a user study with lexicographers gathering information from corpora before and after cleaning. Evaluation along these lines would have made a stronger case for the importance of the thesis work. I also found the evaluation in the de-duplication chapter somewhat unsatisfactory, as no gold standard was used, and only indirect measures of quality were provided. Finally, given that dealing with the fragmentation problem is a particularly innovative aspect of the thesis, I would have liked to see a more extensive and targeted evaluation of this aspect of the algorithms.

UNIVERSITÀ DEGLI STUDI
DI TRENTO

Centro Interdipartimentale Mente/Cervello

The thesis is very well-written (in excellent English) and its contents are generally clear, well-structured, and supported by useful figures and tables. A minor complaint pertains to chapter 3, where the author should have made it more clear that section 3.5 introduces a new algorithm, and he should have provided more general intuitions about the motivations, goals and inner works of the algorithm.

At the defense, I would like the candidate to discuss my main criticism above, in particular suggesting how, in future work, the impact of Web cleaning on the quality of NLP and/or lexicography tasks could be quantified. I would also be interested in how the author intends to extend his work to deal with typical Web 2.0 text, such as the very short and highly connected textual contents of Twitter, YouTube, etc. How are the notions of boilerplate, duplication and fragmentation (and the proposed methods to deal with these problems) affected?

To conclude, Jan Pomikálek has written an interesting and innovative dissertation thesis. He has demonstrated ability to work independently and creatively in computational/corpus-based linguistics and computer science/informatics in general. The thesis meets the standard requirements imposed on a dissertation work in these fields. I strongly recommend to accept the submitted thesis.

Location and date: Rovereto (TN), Italy, July 13th 2011

Sincerely,

Marco Baroni

Corso Bettini, 31 – 38068 Rovereto (TN), Italy - Tel. +39 0464/808612, Fax +39 0464/808690